**Big Data y Machine Learning para Economía Aplicada**

**Taller 2 - 2023-02**

**Link del repositorio en Github:** [**https://github.com/djfarfan10/Taller-2**](https://github.com/djfarfan10/Taller-2)

**Equipo de competencia en Kaggle: *Equipo Rocket***

Angie Ariza Quitian - 201325848

Andrés Felipe Diaz Barreto - 200610686

**Presentado por**

Duvan Javier Farfán López - 201317299

Daniel Eduardo Sandoval - 200712968

1. **Introducción**

De acuerdo con la literatura relacionada con los precios de las viviendas, este se suele determinar a través del modelo de precios hedónicos, que busca estimar el valor de un bien inmueble a través de las características que poseen los mismos y que los hacen diferenciados de otros (Taylor, L. 200X). Es así como surgen variables o características como la cercanía a lugares de servicios (como hospitales, colegios, parques, estaciones de policía, entre otros), y la estructura propia de la construcción que pueden influir en el valor de los inmuebles. Estás características se basan en aquellos atributos propios de los apartamentos y casas del análisis, así como de la ubicación. Según la literatura, dentro de las características propias de un inmueble, los atributos que se destacan son la cantidad de habitaciones, dormitorios y baños que posee, así como la antigüedad de estos, mientras que en cuanto a la ubicación se relacionan aspectos de bienes no mercadeables como la calidad del aire, el ruido, la seguridad, entre otros.

Para este problema set de predicción de precios de vivienda se usaron variables como el área cuadrada de los inmuebles, el estrato la cantidad total de baños y habitaciones, la disponibilidad de parqueaderos propios, depósitos, terrazas y balcones, así como las distancias a Transmilenio, parques, hospitales, colegios, universidades y Comandos de Atención Inmediata – CAI y los delitos registrados durante 2019 por UPZ. Las bases de datos iniciales para realizar el *train* y el *test* contaban con 38.644 y 10.286 observaciones respectivamente, a las cuales se les realizo un ejercicio de limpieza buscando información del texto de la descripción del inmueble eliminando *missing values* que no eran posibles de obtener e imputando valores por la media o la mediana dependiendo de las características de cada variable.

Con la base limpia se estimaron más de 35 modelos, entre los que se usaron los métodos de predicción lineal, Lasso, Ridge, árboles (CART), *Random Forest* y *Boosting*. Se seleccionaron 28 de los anteriores usando la reducción del MAE del *train* como criterio para cargue en Kaggle y se concluyó que la mejor respuesta la daban los Random Forest, seguida por Boosting, Lasso y Ridge. La mejor estimación se obtuvo con este con este tipo de método, arrojando un MAE Test contra los precios de Kaggle de 210.634.496,815 COP usando el 20% de los datos de las predicciones.

1. **Data**

Con el objetivo de predecir los precios de viviendas en Chapinero se usaron dos bases de datos, *train* y *test*, con información asociada a casas y apartamentos en Bogotá, las cuales contenían 38.644 y 10.286 observaciones respectivamente. Al realizar una primera inspección de cada una de las bases se observó que algunas variables importantes para la estimación de precios como el área de los inmuebles, el número de baños y habitaciones contaba con una gran cantidad de *missing values*. También se observó que de la base de *train* no se contaba con la descripción de 9 observaciones, por lo que se eliminaron al no ser de utilidad para la extracción de la información faltante. La distribución de inmuebles en Bogotá para las bases de *train* y *test* se puede ver en la figura 1

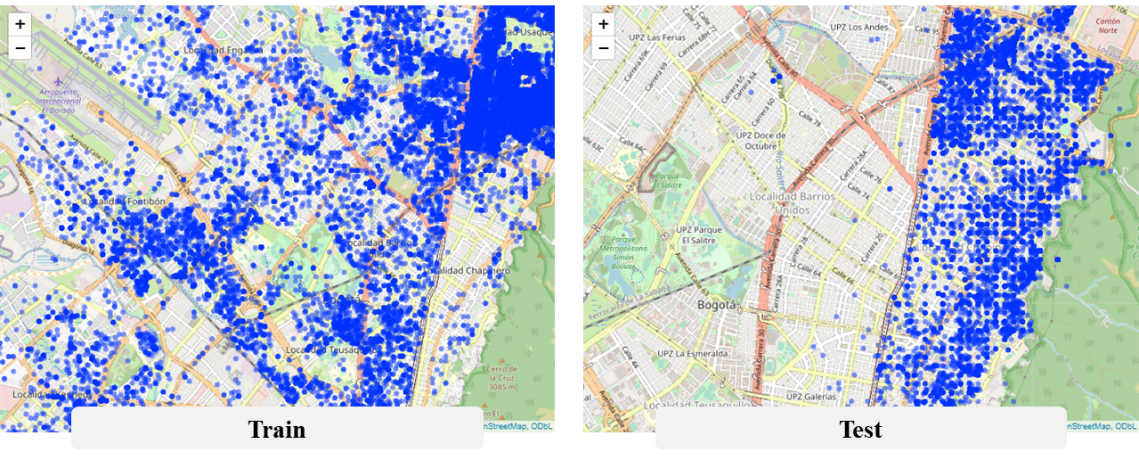


Figura 1. Distribución de inmuebles en Bogotá de las bases de datos train y test para el ejercicio de predicción

Fuente: Elaboración propia con mapas generados en R.

Teniendo en cuenta lo anterior, se extrajo información a partir de la columna *description* para áreas y baños, y se comparó con la información que ya se suministraba. Para cada caso se seleccionó el mayor número de la observación para dichas variables y se consolidó la información en una columna final. Dado que no fue posible extraer la totalidad de información, para los valores de NAN de las variables de área y baños se imputaron estas observaciones faltantes por la mediana de cada variable. El área fue truncada hasta valores máximos de 1.000 metros cuadrados y mínimos de 30 metros cuadrados.

Se realizó también la extracción del tipo de vivienda de la descripción y se comparó con la suministrada en la base, defendiendo como la información final la extraída del texto suministrado en cada base de datos.

Además de las variables previamente mencionadas, usando herramientas de código para extracción de texto se crearon nuevas variables que permitieron identificar las características propias de cada vivienda en la muestra. Las variables creadas fueron:

* *Parqueadero*: Variable dicótoma: 1 si la vivienda tiene parqueadero, 0 si no tiene
* *Baño social*: Variable dicótoma calculada a partir de información de extraída para baños y habitaciones - 1 si la cantidad de baños es mayor a el número de habitaciones, 0 si no lo es
* *Deposito*: Variable dicótoma: 1 si la vivienda cuenta con deposito o garaje, 0 si no tiene
* *Estado de construcción*: Variable dicótoma: 1 si la vivienda es nueva, 0 si no lo es
* *Estado de remodelación*: Variable dicótoma: 1 si la vivienda está remodelada, 0 si no lo es
* *Terraza o balcón*: Variable dicótoma: 1 si la vivienda tiene terraza o balcón, 0 si no cuenta con ninguno de los dos

Los valores que no se pudieron extraer (NAN) para las variables antes mencionadas fueron reemplazados con 0, con la finalidad de no sobreestimar los precios de las viviendas, teniendo en cuenta el objetivo del problema set.

**Variables externas**

En esta sección se describen las variables que fueron tomadas de fuentes externas a las bases de datos proveídas en el ejercicio. La selección de alguna de estas variables se hizo a partir de un estudio anterior (Toloza-Delgado et al., 2021), en el que identifican la importancia que tienen el estrato socioeconómico, la distancia a estaciones de TransMilenio y la distancia a parques.

Por otro lado, de acuerdo con Chacón Rodríguez, V. 2018), los instrumentos de Unidad de Planeamiento Zonal (UPZ) –reemplazadas en el 2022 por las Unidades de Planeamiento Local (UPL)– fueron utilizados para agrupar zonas homogéneas de la ciudad, por lo que pueden ser segmentaciones que capturen parte del precio de la vivienda.

**Distancia a estaciones de TransMilenio**

Las ubicaciones de las estaciones de TransMilenio fueron tomadas de la página principal de la entidad (disponibles en un archivo geojson), teniendo en cuenta que al utilizar el atributo de paradas de buses en la API de OpenStreetMap salían mezcladas algunas cuantas estaciones para buses que no son BRT. En total, se tienen 149 estaciones de TransMilenio distribuidas en 12 troncales del sistema.

**Estratos socioeconómicos**

En la plataforma distrital de Datos Abiertos, se tiene, entre otros, las manzanas georreferenciadas según su estrato socioeconómico. Este atributo es asignado por las alcaldías para clasificar a una vivienda según características físicas externas e internas, su entorno inmediato y su contexto habitacional o funcional (DANE, 2023).

Por la georreferenciación inadecuada de algunas viviendas y porque hay áreas de la ciudad que no están incluidas dentro de estas manzanas (vías, áreas públicas, zonas comerciales), se asignó el estrato de la vivienda según su cercanía al centroide del polígono1, es decir, la vivienda tenía el valor del estrato del polígono cuyo centroide estuviera más cerca. En la siguiente figura se muestra el resultado de la asignación:

1Inicialmente se realizó la asignación con el criterio de que la vivienda estuviera dentro del polígono, sin embargo, alrededor del 30% de los datos no estaban dentro de las manzanas delimitadas. Para ilustrar véase los puntos señalados en la Figura 2, que están parques, zonas verdes o vías principales.

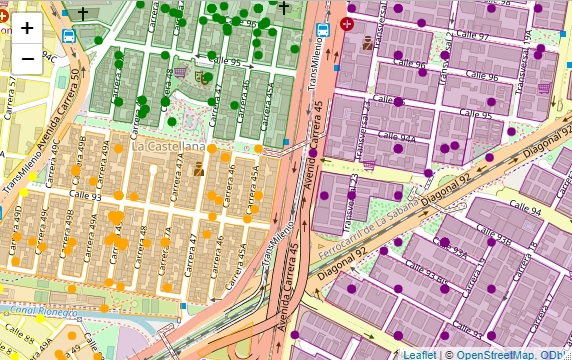


Figura 2. Muestra de asignación de los estratos socioeconómicos a las viviendas. Los puntos señalados en los recuadros rojos son ejemplos de la georreferenciación inadecuada, lo que resulta en que haya viviendas por fuera de las manzanas residenciales.

Fuente: Elaboración propia con mapas generados en R.

La zona de la anterior figura es muy interesante porque en un radio de 1.000 metros se pueden ver cuatro estratos socioeconómicos: tres (amarillo), cuatro (naranja), cinco (verde) y seis (purpura). Los puntos representan las viviendas de las bases de datos y los polígonos sombreados son los que estaban disponibles en Datos Abiertos

**Unidades de Planeamiento Zonal y delitos por UPZ**

Los polígonos de la UPZ fueron tomados del Laboratorio Urbano Bogotá (Laboratorio Urbano, 2012) y emparejados con las viviendas de las bases de datos según su ubicación espacial. En total, se tienen 112 UPZ correspondientes 19 localidades dentro de la ciudad. Adicional a lo anterior, se buscó información de delitos en Bogotá y se encontró que estaba relacionada a las UPZ, por lo que se generó una variable que consolidaba la totalidad de delitos por cada una y se asoció a los inmuebles. Para este ejercicio se usaron datos de 2019, dado que los valores de 2020 a 2022 pueden estar sesgados por el efecto de la pandemia del COVID-19 y aún no se contaba con la totalidad de información para 2023. Los delitos incluían robo a personas, robo a comercios, homicidios, delitos sexuales, hurto de celulares, violencia intrafamiliar, entre otros.

**Hospitales, CAIs, universidades, colegios y parques en Bogotá**

Así mismo, revisando las categorías contenidas en *leisure* y *amenity* de Open Street Maps, y de acuerdo con la literatura revisada, se obtuvo información de parques, hospitales, universidades, colegios, comandos de atención inmediata (CAIs), con la cual se determinó la distancia mínima de cada inmueble a cada uno de estos espacios. En la figura XX se observa la distribución de las categorías usadas a lo largo de Bogotá.

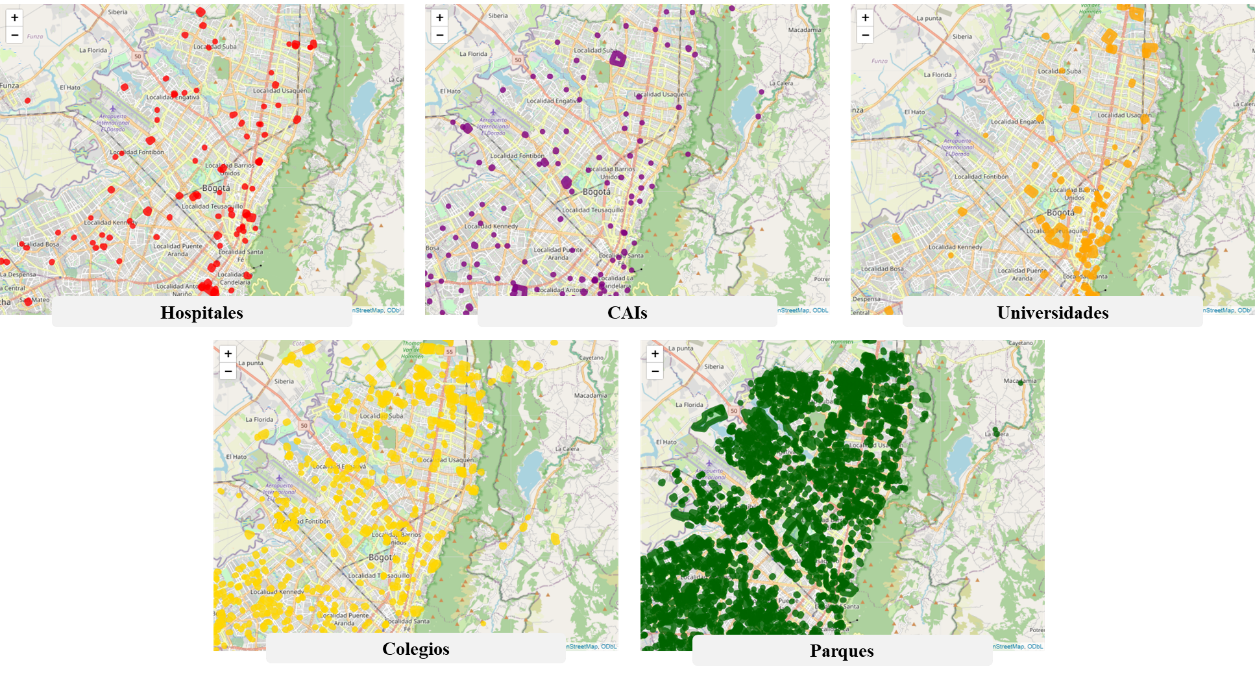


Figura 3. Distribución de hospitales, CAIs, universidades, colegios y parques en Bogotá

Fuente: Elaboración propia con mapas generados en R.

Una vez extraídas las variables de texto y las variables externas se realizó la unión de las mismas para generar las bases definitivas mediante funciones de *merge* usando como llave la variable de *property\_id*. Se revisó el tipo de variables y se ajustó el tipo de las categóricas.

Incluir un análisis descriptivo de los datos. Como mínimo, debes incluir una tabla de estadísticas descriptivas y dos mapas con su interpretación. Sin embargo, se espera un análisis profundo que ayude al lector a comprender los datos, su variación y la justificación de tus elecciones de datos. Utiliza tu conocimiento profesional para agregar valor a esta sección. No lo presentes como una "lista seca" de ingredientes.

1. **Modelo y resultados**

Las variables utilizadas para entrenar el modelo se encuentran en la siguiente tabla, con una breve descripción de cada una de ellas:

Tabla 1. Variables del modelo escogido

| Tipo de atributo | | Atributo | Explicación |
| --- | --- | --- | --- |
| Físico | Área construida de la vivienda | | Variable continua que describe el tamaño en metros cuadrados de la vivienda |
| Físico | Tipo de propiedad | | Variable categórica con dos clases: apartamento o casa |
| Físico | Cantidad de habitaciones | | Variable discreta con la cantidad de habitaciones que tiene la vivienda |
| Físico | Cantidad de baños | | Variable discreta con la cantidad de baños que tiene la vivienda |
| Físico | Existencia de terraza o balcón | | Variable categórica con dos clases: tiene terraza o no tiene terraza |
| Físico | Existencia de parqueadero | | Variable categórica con dos clases: tiene terraza o no tiene terraza |
| Físico | Existencia de depósito | | Variable categórica con dos clases: tiene depósito o no tiene depósito |
| Ordenamiento | UPZ | | Variable categórica con 112 clases, correspondientes a las UPZ existentes al 2019 |
| Servicios | Distancia la estación de TransMilenio | | Variable continua que describe la distancia a la estación de TransMilenio más cercana. Polinomio orden 2 |
| Servicios | Distancia al parque | | Variable continua que describe la distancia al parque más cercano. Polinomio orden 2 |
| Servicios | Distancia a la universidad | | Variable continua que describe la distancia a la universidad más cercana. Polinomio orden 2 |
| Servicios | Distancia al hospital | | Variable continua que describe la distancia al hospital más cercano. Polinomio orden 2 |
| Servicios | Distancia al colegio | | Variable continua que describe la distancia al colegio más cercano. Polinomio orden 2 |
| Social/Económico | Estrato | | Variable categórica con 6 clases, correspondientes a los estratos socioeconómicos en Colombia |
| Social/Económico | Cantidad de delitos en el 2019 en la UPZ | | Variable continua que describe la cantidad de delitos cometidos en el 2019 por UPZ, descritos en la sección anterior |
| Social/Económico | Distancia al CAI | | Variable continua que describe la distancia al Comando de Atención Inmediata (CAI). Polinomio orden 2 |

Nota: Las distancias estimadas están en metros y dentro del entrenamiento del modelo fueron incluidas como un polinomio de orden 2, luego de ver que, a través de varias modificaciones, el modelo tenía un mejor desempeño con esta especificación para todas las distancias.

Con estas variables el entrenamiento del modelo fue realizado a través de *Random Forest,* que es un método basado en árboles, en el que construyen una cantidad de árboles de decisión en muestras de entrenamiento que son generadas a través de un *Bootstrap*. Un parámetro crítico en este método es la cantidad de variables que son consideradas en cada división de los respectivos árboles, denominado como o en el código como .

Inicialmente, se definió el parámetro con base en la literatura [YYY1] teoría, en la que se indica que típicamente se define como la raíz cuadrada de la cantidad total de predictores. Para este caso, teniendo en cuenta la cantidad total de predictores (110), se definió inicialmente el como 10. Sin embargo, al realizar varias iteraciones con ajustes en el , se evidenció que la cantidad con el menor error absoluto (MAE, por sus siglas en inglés) era de 30.

Otro parámetro importante es la cantidad de dobleces o *folds* que se utilizan para la división de la base de datos de muestra para entrenar el modelo. A través de un proceso iterativo que comenzó con cinco *folds* y terminó con veinte *folds*, se identificó que cinco era la mejor cantidad en términos de desempeño y esfuerzo computacional (con diez y veinte *folds* las mejoras en MAE eran menos de 3 millones de pesos, lo que representaba una mejora de menos del 3% respecto al MAE del mejor modelo).

El último parámetro que fue tenido en cuenta para este proceso iterativo fue el tamaño del nodo o de la hoja de los árboles generados en el proceso. Para evitar el sobreajuste del modelo o un truncamiento forzado de la profundidad de los árboles, se evaluó el resultado con dos valores: 3 y 5.

La siguiente tabla muestra algunos de las varias alternativas estudiadas para escoger el mejor modelo dentro del método de *Random Forest* –que era el que tenía mejor desempeño con la base de datos de entrenamiento y con la validación en Kaggle (Las predicc):

Tabla 2. Alternativas evaluadas con Random Forest

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| mtry | Folds | Tamaño de nodos | MAE - Train | Diferencia porcentual con el mejor modelo |
| 30 | 10 | 5 | 96.776.686 | 0% |
| 20 | 10 | 3 | 97.915.701 | 1% |
| 20 | 10 | 5 | 99.325.911 | 3% |
| 10 | 20 | 5 | 118.459.734 | 18% |
| 10 | 10 | 5 | 119.370.083 | 19% |
| 10 | 5 | 5 | 121.089.040 | 20% |

Como se puede ver con esta muestra de opciones estudiadas, el parámetro más importante era *.* Adicionalmente, ya después de un valor de 30 el esfuerzo computacional era considerable y tomaba varias horas para que se identificara el mejor modelo2.

2 Se realizó un modelo con un parámetro de mtry de 50 y los cambios en el MAE fueron menos del 5%

El mejor modelo fue escogido de 28 opciones que fueron sometidas a validación en Kaggle y varias más que no fueron subidas, pero cuyo MAE en la base de datos entrenamiento fue el criterio para no ser subidos a la competencia. Dentro de estos intentos, se incluyeron los métodos vistos en clase, desde regresión lineal hasta *Boosting,* para ver cuál se comportaba mejor. La siguiente gráfica muestra el MAE para la base de datos de entrenamiento (*Train MAE*) y el resultado en Kaggle cuando se validaba (*Test MAE*) de 8 predicciones, incluyendo el mejor modelo obtenido:

Figura 4. Comparación de resultados obtenidos para 8 estimaciones con diferentes modelos y métodos

Fuente: Elaboración propia con información obtenida de predicciones en R y resultados de MAE en Kaggle

Es importante notar que el resultado del mejor método también obedece a que en este fue el único que se incluyeron otras variables de servicios que se tenían alrededor de la vivienda (distancias a CAI, hospital, universidad, colegio, parque). Si se actualizarán los demás modelos con la especificación más completa se esperaría que los resultados mejoraran, pero no que el orden se alterara ya que antes de que se incluyeras estas variables, el método de *Random Forest* era el que tenía menor MAE3.

3 Con los métodos de regularización se hizo la inclusión de estas nuevas variables y sus MAE se redujeron en menos de 1 millón COP, es decir, menos del 1%.

1. Conclusiones y recomendaciones

Con la base limpia se estimaron más de 35 modelos, entre los que se usaron los métodos de predicción lineal, Lasso, Ridge, árboles (CART), Random Forest y Boosting. El cálculo y verificación de reducción del MAE para el train permitió seleccionar seleccionaron 28 modelos que fueron usados para cargar en Kaggle y estimar el MAE del test. Se concluyó que la mejor respuesta se obtenía de los modelos que usaban los Random Forest, seguida por Boosting, Lasso y Ridge.

La mejor estimación se obtuvo con este con este tipo de método, arrojando que el mejor MAE Test contra los precios de Kaggle fue de 210.634.496,815 COP. Para la mejor predicción obtenida se usaron variables de atributos propios de los inmuebles y características de su ubicación como cercanía a estaciones de Transmilenio, colegios, CAIs y hospitales, lo que soporta la literatura revisada para precios de vivienda a través de precios hedónicos y bienes no mercadeables como la seguridad, el aire y espacios de recreación y de servicios como la atención médica y la educación.

Los modelos lineales fueron los que presentaron la predicción menos acertada y los MAE más altos, sin embargo, se debe agregar que la mejora de estos resultados depende del nivel de limpieza de la base de datos y de la calidad de la información, pues al usar información asociada a distancias los modelos mejoraron sustancialmente.

Para la estimación de modelos de predicción de vivienda, en espacial para Chapinero y entendiendo la dinámica de la zona, se recomienda incluir variables asociadas al ocio, la comida y la vida nocturna como bares, gimnasios, restaurantes, y centros comerciales, que no fue posible incluir en este problema set dada la limitación de información que se encontró en fuentes oficiales como Datos Abiertos, Alcaldía de Bogotá e incluso Open Street Maps.

También se recomienda realizar un análisis de sensibilidad sobre los cambios en las estimaciones con bases de datos que imputen y que no imputen valores para observaciones con NAN. Aunque en este trabajo se realizó lo propio para la variable de área, observando un cambio en los resultados de aproximadamente 5 millones, es importante revisarlo para las demás variables usadas, pues puede significar una mejora frente a las predicciones finales y los MAEs obtenidos.

Bibliografía

[XXX1] Toloza, J., Melo-Marinez, O., Azcarate-Romero, J. (2021). Determinantes del precio de la vivienda nueva en Bogotá para el año 2019: una aproximación a través de un modelo semiparamétrico de regresión espacial. doi:10.17230/ingciencia.17.34.2

[XXX2] Chacón, V. (2018). ZONAS HOMOGENEAS ECONOMICAS – ESTUDIO DE VALOR DE M2 ÁREA PRIVADA CONSTRUIDA DE EDIFICACIONES SOMETIDAS A PROPIEDAD HORIZONTAL EN LA UPZ No. 101 TEUSAQUILLO EN LA LOCALIDAD DE TEUSAQUILLO - BOGOTÁ D.C. <https://repository.udistrital.edu.co/bitstream/handle/11349/14505/ChaconRodriguezVictorDavid2018.pdf?sequence=1>

[XXX3] DANE (25 de octubre del 2023). Preguntas frecuentes estratificación. <https://www.dane.gov.co/files/geoestadistica/Preguntas_frecuentes_estratificacion.pdf>